Geo-spatial Information Science 11(3):191-196
DOI 10.1007/s11806-008-0094-8
Volume 11, Issue 3
September 2008

**Article ID:** 1009-5020(2008)03-191-06
**Document code:** A

# Fuzzy Clustering Validity for Spatial Data

**HU Chunchun MENG Lingkui SHI Wenzhong**

Abstract The validity measurement of fuzzy clustering is a key problem. If clustering is formed, it needs a kind of machine to verify its validity. To make mining more accountable, comprehensible and with a usable spatial pattern, it is necessary to first detect whether the data set has a clustered structure or not before clustering. This paper discusses a detection method for clustered patterns and a fuzzy clustering algorithm, and studies the validity function of the result produced by fuzzy clustering based on two aspects, which reflect the uncertainty of classification during fuzzy partition and spatial location features of spatial data, and proposes a new validity function of fuzzy clustering for spatial data. The experimental result indicates that the new validity function can accurately measure the validity of the results of fuzzy clustering. Especially, for the result of fuzzy clustering of spatial data, it is robust and its classification result is better when compared to other indices.

Keywords fuzzy clustering; spatial data; validity; uncertainty
CLC number P208

## Introduction

Clustering is one of the most useful tools in data mining, and can identify data groups and interesting distribution patterns. Spatial clustering is about partitioning a given spatial data set into groups (clusters) such that the spatial objects within a cluster are more similar to each other than objects in different clusters. Generally, the clusters are non-overlapping and this kind of partitioning is called crisp clustering. However, relations between data may be cut by crisp clustering. Further, the kind of clustering can not express the uncertainty of the kind of data. Especially, clustering is not fit for spatial data because spatial data belong to the complicated data type and have spatial relations. The issue of uncertainty support leads to the introduction of algorithms with fuzzy logic concepts in clustering procedures.

Fuzzy clustering can partition the given data set into groups without being concerned with the data structure in the data set. Yet how do we validate if the partition result is right? One of the most important issues in clustering analysis is the evaluation of clustering, which results in the discovery of the partition that best fits the underlying data. The choice of optimal cluster number and evaluating the cluster results are called clustering validation problems. The issues of clustering validity are focused on two categories of fuzzy validity indices. The first category uses only the membership values, $u_{ij}$, of a fuzzy partition of the data set. The partition coefficient (PC)[1] and partition entropy (PE)[2] proposed by Bezdeck belong to this category. The latter involves both the $\boldsymbol{U}$ matrix and the dataset itself. Indices in this category include the XB index[3] and the FS index[4]. However, many validity indices are limited to focusing only on the distance between cluster centroids when considering the

Received on July 10, 2008.
HU Chunchun, School of Geodesy and Geomatics, Wuhan University, 129 Luoyu Road, Wuhan 430079, China.
E-mail:hccain@yahoo.com.cn

geometry structure in the datum[5], and do not consider the partition quality. Actually, class uncertainty goes with fuzzy clustering, and the degree of uncertainty determines the reliability of the fuzzy partitioning results. In this paper, a new evaluation method of spatial fuzzy clustering is proposed with respect to the uncertainty factor and the spatial features of spatial data.

The remainder of this paper is organized as follows. Section 2 introduces the detection method for the clustered pattern, FCM algorithm and previous validity indices. The new validity index is proposed in Section 3. Section 4 gives the results of the experiments, and Section 5 presents the conclusions.

# 1 FCM algorithm and cluster validity indices

Fuzzy clustering exploits fuzzy techniques for cluster data, and an object can be classified into more than one cluster. To make mining more accountable, comprehensible and with a usable spatial pattern, it is necessary to detect whether the data set has a clustered structure or not before clustering.

## 1.1 Detecting clustered structures

Spatial data have complicated spatial features, spatial autocorrelation and spatial variability. Detecting a clustered structure from a spatial data set can be implemented by statistical analysis. The function test based on distance between points, Quadrat counts, statistical indicators and variograms used to judge spatial autocorrelation, are all typical detection methods. Especially, the function test based on distance between points is easy and intuitive, and it is broadly applied in many fields. The *K*-function and the *L*-function belong to this category. The *L*-function is another expression of the *K*-function. Its product curve is easier to judge compared with that of the *K*-function. The value of the *L*-function which is greater than zero indicates that there are clustered structures in the data set.

For example, the detected result is shown in Fig.2 by using the *L*-function for the data set shown in Fig.1. In Fig.2, the real line denotes the variable values of the *L*-function corresponding with distance scale *d*, and the dashed line is the line *x*=0. From this figure, we can see that this data set has a clustered structure at the range of (0, 4). So we can judge the clustered pattern by exploiting these detection methods.

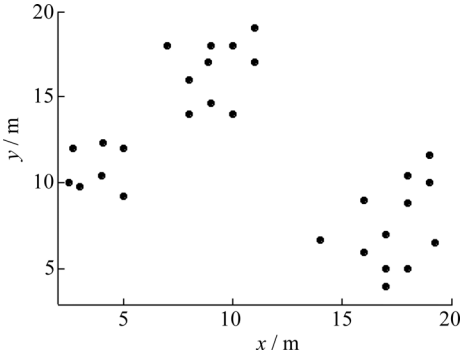


**Fig.1 Data set**

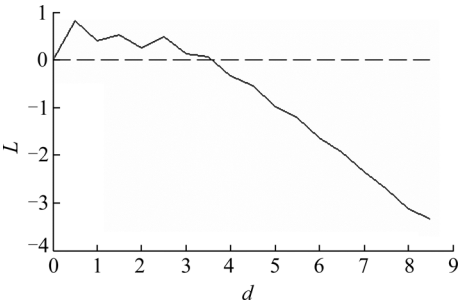


**Fig.2 Detected result**

## 1.2 The FCM algorithm

A common fuzzy clustering algorithm is the Fuzzy *c*-Means (FCM), an extension of classical *c*-Means algorithm for fuzzy applications[6]. FCM attempts to find the most characteristic point in each cluster, which can be considered as the "center" of the cluster, and the membership grade of each object in the clusters. The FCM clustering algorithm has been widely used to obtain the fuzzy *c*-partition. However, the algorithm may fall into local optima due to the initialization of cluster centroids[5]. It is a kind of fuzzy clustering algorithm-based object function. Given a dataset $\boldsymbol{X}=\{X_1, X_2, \cdots, X_n\}$ with dimension *s*, the object of FCM is to partition dataset $\boldsymbol{X}$ into *c* homogeneous fuzzy clusters by minimizing the function $J_m$

$$J_m = \sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij})^m d^2(X_j, V_i) \quad (1)$$

where *c* is the number of clusters; *n* is the number of data, and $u_{ij}$ is the membership degree of data point

$X_j$ belonging to the fuzzy cluster $C_i$; $V_i$ is the $i^{\text{th}}$ cluster centroid; $m$ is the weighting exponent and controls the fuzziness of membership of each datum[5]; $d(X_j, C_i)$ represents the Euclidean distance between $X_j$ and $V_i$.

The execution of function $J_m$ is an optimization problem, and approximate optimization of $J_m$ is based on iteration through computing the membership degree $u_{ij}$ and cluster centroid $V_i$ for its local extrema. The limited condition of function $J_m$ is that the sum of membership degree ($u_{ij}$), with $X_j$ belonging to each of cluster $C_i$, equals 1:

$$\sum_{i=1}^{c} u_{ij} = 1, \quad 1 \leqslant j \leqslant n; \quad 0 \leqslant u_{ij} \leqslant 1 \tag{2}$$

The FCM algorithm is carried out in the following steps.

**Step 1** Initialize threshold $\varepsilon$ and cluster centroids $V^{(0)}$, set $k = 0$.

**Step 2** Give a predefined number of cluster $c$ and a chosen value of $m$.

**Step 3** Compute the matrix of the membership degree $U(k) = [u_{ij}]$ for $i$=1,2,…,$c$ using

$$u_{ij} = \frac{(d^2(X_j, V_i))^{\left(-\frac{1}{m-1}\right)}}{\sum_{i=1}^{c} (d^2(X_j, V_i))^{\left(-\frac{1}{m-1}\right)}} \tag{3}$$

**Step 4** Update the fuzzy cluster centroid $V_i(k+1)$ for $i - 1, 2, \cdots, c$ using

$$V_i(k+1) = \frac{\sum_{j=1}^{n} u_{ij}^{m}(k) X_j}{\sum_{j=1}^{n} u_{ij}^{m}(k)} \tag{4}$$

**Step 5** If $\| V(k) - V(k+1) \| < \varepsilon$, then iteration is halted; otherwise return to step 3.

The FCM algorithm always converges on a local maximum value through the above iteration calculation[2].

### 1.3 Validity index

As mentioned above, the validity indices are mainly divided into two categories according to their properties: involving only the membership values or involving the membership values and the dataset. Bezdek's partition coefficient (PC) and partition entropy (PE) belong to the first category, defined as

$$PC = \frac{\sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik}^2}{n} \tag{5}$$

$$PE = -\frac{1}{n} \left\{ \sum_{k=1}^{n} \sum_{i=1}^{c} \left[ u_{ik} \log_a (u_{ik}) \right] \right\}, \quad a \in (1, \infty) \tag{6}$$

where $c$ and $n$ are the number of clusters and data respectively, and $u_{ij}$ is the membership degree of data point $X_j$ in the fuzzy cluster $C_i$. Eq.(5) indicates that an optimal value of $c$ is obtained by maximizing $v_{PC}$, for its value increases with an increase in $u_{ij}$. Thus, the cluster with a high value of PC represents a compact cluster well. Contrarily, an optimal partition is obtained by minimizing PE. However, the two validity indices only consider the membership degree of data so that they lack a direct connection to the geometry of the data[7].

The Xie-Beni index belongs to the other category, which involves the membership values and the dataset itself. It is a compact and separate fuzzy validity function[3] and is defined as

$$v_{XB} = \frac{\sum_{j=1}^{n} \sum_{i=1}^{c} u_{ij}^2 \| x_j - v_i \|^2}{n \cdot \min\limits_{i \neq k} \| v_i - v_k \|^2} \tag{7}$$

Eq.(7) is explained as the ratio of total compactness to the separation of the fuzzy $c$-partition. For compact and well-separated clusters, small values of $v_{XB}$ are expected. Therefore, the optimal cluster $c$ is obtained by finding the fuzzy $c$-partition with the smallest value of $v_{XB}$.

## 2 New validity index

For the problem of clustering validation, the objective of most methods is to seek clustering patterns in which most of the data in the dataset exhibit a high degree of membership in one cluster. The validation indices PC and PE belong to this category. The result of research focused on the two indices indicate that the index PC always tends to descend with an increase in the cluster number $c$. The value of index PE tends to increase with cluster number $c$ as well. The two indices only consider degree of membership, not the geometry structure of the data set itself. At the same time, they ignore the uncertainty factor in the fuzzy partitioning process. So the

partition results of the two indices are not perfect enough and can not exactly evaluate the validity of spatial fuzzy clustering.

### 2.1 Important factors of impact on validity evaluation

The factors involved can involve various contents for the validity problem of spatial fuzzy clustering. The uncertainty factor in fuzzy partition and spatial features of spatial data are key parts except for the degree of membership and the data set itself.

For the fuzzy partition of spatial data having uncertainty, the class uncertainty is an important factor in validity evaluation. A successful class schema should include the amount of significant information[8]. The optimal results of fuzzy partition should be that any cluster partitioned include amounts of significant information. The data set giving support to each cluster is expressed as $1/N\cdot(\sum_{i=1}^{N}u_{ij}{}^{q})$ in information theory, and the support can reflect the magnitude of amount of information in this cluster. If the support is higher, then the cluster includes more information. Moreover, the clustering results are more reliable in fuzzy partition. The $-1/N\cdot\left[\sum_{i=1}^{N}\log_2(u_{ij})\right]$[8] can be used to denote the degree of cluster uncertainty, and it evaluates the degree of cluster uncertainty according to the degree of membership in a cluster. The higher the degree of uncertainty is, the fuzzier the partition is. The $\log_2(n_c)$ can denote the highest degree of uncertainty, where $n_c$ is the cluster number of fuzzy partition. The deviation between the degree of uncertainty of a cluster and its maximal value can reflect the reliability of fuzzy partition results. The greater the deviation is, the more reliable the partition results are.

For spatial fuzzy clustering, the spatial features of spatial data, which include intra-cluster comparability and inter-cluster difference, can reflect the partition results of fuzzy clustering. The intra-cluster comparability reflects the measurement of separation between each spatial object and the cluster centre within a cluster. On the other hand, the inter-cluster difference shows the measurement of separation between cluster centres. The ratio of inter-clusters to intra-clusters can reflect the partition results. A higher ratio means a larger separation between inter-clusters and smaller compactness within a cluster.

### 2.2 Definition of validity function

To take into account the two important factors above, a new validity function of spatial fuzzy clustering is defined as follows:

$$IFV=\frac{1}{n_c}\sum_{j=1}^{n_c}\{\frac{1}{N}\sum_{i=1}^{N}u_{ij}{}^{2}\cdot(\log_2(n_c)-\frac{1}{N}\sum_{i=1}^{N}\log_2(u_{ij}))^2\}\cdot\frac{SD_{\max}}{\overline{\sigma_D}} \tag{8}$$

In Eq.(8), the $(\log_2(n_c)-\frac{1}{N}\sum_{i=1}^{N}\log_2(u_{ij}))^2$ denotes the uncertainty degree of the $j^{\text{th}}$ cluster, and the larger value means a smaller uncertainty degree and more reliable partition results. The $SD_{\max}=\max_{k\neq j}\|V_k-V_j\|^2$ denotes the maximal distance, while $\overline{\sigma_D}=\frac{1}{n_c}\sum_{i=1}^{n_c}(\frac{1}{N}\sum_{i=1}^{N}\|X_i-V_j\|^2)$ expresses the even deviation between each object and the cluster centre. A larger ratio of $SD_{\max}\sqrt{\sigma_D}$ denotes better cluster results.

To sum it up, if we find one or more optimal partitions of the data set $X$ for each $c=2,3,\cdots,n_c$, and satisfy

$$IFV(U^*,C^*;X)=\max_{\Omega_c}IFV(U,C;X) \tag{9}$$

then the value of *IFV* is said to yield the most optimal fuzzy *c*-partition of data set ***X***.

## 3 Experimental results

In this section, we evaluate the effectiveness of the proposed index IFV on three kinds of data sets, and it is compared with two validity indices: Bezdek's partition coefficient (PC) and partition entropy (PE).

For obtaining reliable experimental results, the parameters of the FCM are set to the termination criterion $\varepsilon=0.001$ and weighting exponents $m=1.5$, $m=2$ and $m=2.5$. The choice of cluster number $c$ is decided by the size of the data set.

### 3.1 Experiment on a non-overlapping data set

In this experiment, a more regular 2-D test data set is evaluated. The data set includes 55 points without overlapping. We made several runs of the FCM algo-

rithm with each value of $c$=2,3,⋯,8. The experimental results are listed in Table 1. The optimal clustering number $c$ chosen by each index is highlighted in bold in Table 1. From Table 1, we see that three indices correctly discover the optimal $c$=4. The result is coincident with the actual classification of the data set. Hence, for the kind of data set without overlapping, the validity indices PC, PE and the proposed index IFV can evaluate clustering validity, and the partition results are perfect.

**Table 1 The experimental results of the test data**

| | $m$=1.5 | | | $m$=2 | | | $m$=2.5 | | |
|---|---|---|---|---|---|---|---|---|---|
| $c$ | PC | PE | **IFV** | PC | PE | **IFV** | PC | PE | **IFV** |
| 2 | 0.70 | 0.66 | 0.11 | 0.70 | 0.98 | 0.12 | 0.70 | 0.66 | 0.13 |
| 3 | 0.69 | 0.77 | 1.26 | 0.69 | 0.72 | 1.50 | 0.67 | 0.73 | 1.37 |
| **4** | **0.81** | **0.60** | **3.24** | **0.80** | **0.59** | **3.29** | **0.81** | **0.59** | **3.29** |
| 5 | 0.74 | 0.78 | 2.91 | 0.74 | 0.78 | 2.97 | 0.74 | 0.85 | 2.27 |
| 6 | 0.70 | 0.92 | 2.76 | 0.70 | 1.03 | 2.05 | 0.70 | 0.92 | 2.74 |
| 7 | 0.64 | 1.03 | 2.48 | 0.65 | 1.04 | 2.17 | 0.64 | 1.17 | 1.88 |
| 8 | 0.62 | 1.16 | 1.92 | 0.61 | 1.18 | 2.05 | 0.62 | 1.17 | 2.05 |

### 3.2 Experiment on an overlapping data set

A real dataset IRIS is evaluated in this experiment. The data set has three physical groups and each group has 50 data with four dimensions. Two of the three groups are overlapping, while the third is separated from the other two groups. From the experimental results listed in the Table 2, the proposed index points to the cluster number $c$=3, while other indices point to cluster number $c$=2. Both $c$=2 or $c$=3 for IRIS are the optimal choices[9]. So the three indices can exactly evaluate fuzzy clustering validity for this kind of data set with overlapping.

**Table 2 The experimental result of IRIS data**

| | $m$=1.5 | | | $m$=2 | | | $m$=2.5 | | |
|---|---|---|---|---|---|---|---|---|---|
| $c$ | PC | PE | **FV** | PC | PE | **IFV** | PC | PE | **IFV** |
| **2** | **0.89** | **0.28** | 1.84 | **0.89** | **0.28** | 1.83 | **0.89** | **0.28** | 1.85 |
| **3** | 0.77 | 0.58 | **2.49** | 0.77 | 0.58 | **2.50** | 0.78 | 0.57 | **2.55** |
| 4 | 0.70 | 0.81 | 2.23 | 0.68 | 0.84 | 1.66 | 0.70 | 0.81 | 2.24 |
| 5 | 0.62 | 1.06 | 1.65 | 0.65 | 1.01 | 1.80 | 0.64 | 1.03 | 1.60 |
| 6 | 0.57 | 1.19 | 1.55 | 0.60 | 1.18 | 1.44 | 0.57 | 1.20 | 1.55 |
| 7 | 0.54 | 1.33 | 1.38 | 0.53 | 1.36 | 1.40 | 0.53 | 1.35 | 1.40 |
| 8 | 0.50 | 1.48 | 1.18 | 0.54 | 1.47 | 1.08 | 0.52 | 1.52 | 1.11 |

### 3.3 Experiment on spatial data set

The data set used to evaluate the validity of the proposed index IFV is a spatial data set named NUMP[10] in this experiment. It includes 4 495 spatial points with longitude and latitude coordinates. We implemented the experiment on the spatial data set with three indices PC, PE and IFV. The experimental results are listed in Table 3. From Table 3, we can see that index IFV yields the maximal value when cluster number $c$=7 and the weighting exponent $m$=1.5 and $m$=2, while indices PC and PE achieve the maximal value 0.78 and the minimal value 0.52.

**Table 3 The experimental results of spatial data**

| | $m$=1.5 | | | $m$=2 | | | $m$=2.5 | | |
|---|---|---|---|---|---|---|---|---|---|
| $c$ | PC | PE | **IFV** | PC | PE | **IFV** | PC | PE | **IFV** |
| 2 | **0.78** | **0.52** | 0.39 | **0.78** | **0.52** | 0.39 | **0.78** | **0.52** | 0.39 |
| 3 | 0.63 | 0.90 | 0.64 | 0.63 | 0.90 | 0.64 | 0.63 | 0.90 | 0.64 |
| 4 | 0.61 | 1.03 | 0.99 | 0.61 | 1.03 | 0.99 | 0.61 | 1.03 | 0.99 |
| 5 | 0.55 | 1.27 | 1.33 | 0.59 | 1.17 | 1.45 | 0.59 | 1.18 | 1.34 |
| 6 | 0.58 | 1.26 | 1.21 | 0.58 | 1.26 | 1.21 | 0.58 | 1.26 | 1.20 |
| **7** | 0.51 | 1.29 | **1.72** | 0.59 | 1.29 | **1.69** | 0.59 | 1.29 | **1.72** |
| 8 | 0.51 | 1.41 | 1.26 | 0.58 | 1.37 | 1.51 | 0.59 | 1.33 | **1.81** |
| 9 | 0.58 | 1.41 | 1.49 | 0.57 | 1.41 | 1.48 | 0.57 | 1.41 | 1.42 |
| 10 | 0.60 | 1.38 | 1.59 | 0.58 | 1.42 | 1.46 | 0.58 | 1.42 | 1.47 |

In order to compare the proposed index IFV with the other indices PC and PE, we describe the relation between cluster number $c$ and the values of the three indices yielded from the experiment by the curve shown in Fig.3. The values of PC and PE have a tendency of increasing with cluster number $c$, and the highest point of the PC curve and the lowest point of the PE curve are located at $c$=2 in Fig.3. Their optimal cluster number is 2. The IFV curve tends to increase with cluster number $c$ as well, but it has a wave crest at cluster number $c$=7, and its optimal

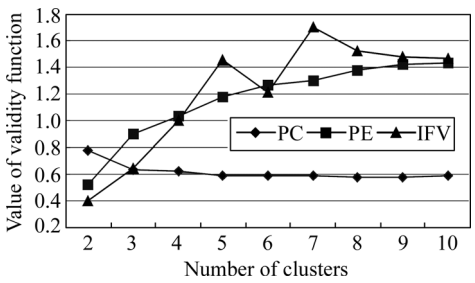


**Fig.3 The validity of clustering for spatial data (NUMP)**

cluster number is 7. The partition results at $c=2$ and $c=7$ are shown in Fig.4 and Fig.5 separately, and each cluster is expressed by a different shape and color. The spatial distribution is not enough, as is obvious in Fig.4. In contrast, the seven clusters in Fig.5 can exactly reflect the clustered tendency.

Hence, indices PC and PE can not discover the correct cluster number for the spatial data set named NUMP in this experiment, but the proposed IFV can not only identify the optimal cluster number, it can also yield the perfect figure of the partition results.

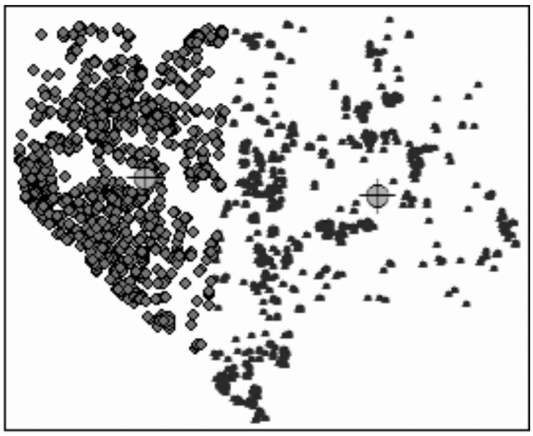

**Fig.4 Fuzzy classification of spatial data at $c=2$**

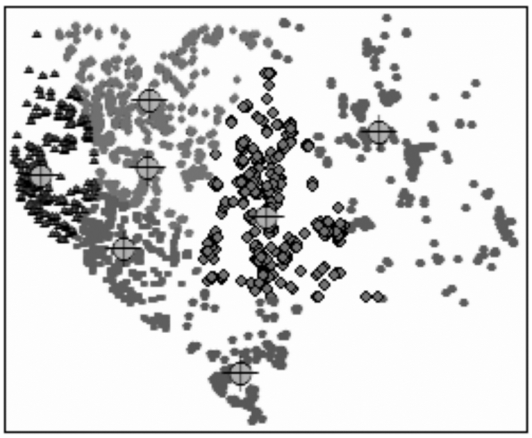

**Fig.5 Fuzzy classification of spatial data at $c=7$**

## 4 Conclusions

Validity evaluation of fuzzy clustering is a key problem in whether fuzzy clustering can be exploited successfully or not. This paper proposes a validity index IFV for spatial fuzzy clustering based on an uncertainty factor in the fuzzy partition process and spatial location features of spatial data. The experimental results indicate that the index can identify the correct cluster number for data sets with overlapping and non-overlapping structures. Obviously, its veracity is higher than that for PC and PE. Especially, for validity evaluation of spatial fuzzy clustering, the partition results yielded by the index IFV are better compared with PC and PE.

## References

[1] Bezdek J C (1980) A convergence theorem for the fuzzy ISODATA clustering algorithm[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2): 1-8

[2] Bezdek J C (1981) Pattern recognition with fuzzy objective function algorithms[M]. New York: Plenum Press

[3] Xie X L, Beni G (1991) A validity measure for fuzzy clustering[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8): 841-847

[4] Fukuyama Y, Sugeno M (1989) A new method of choosing the number of clusters for the fuzzy c-means method[C]. Proceedings of the Fifth Systems Symposium , Japanese

[5] Kim D-W, Lee K H, Lee D (2003) Fuzzy cluster validation index based on inter-cluster proximity[J]. *Pattern Recognition Letters*, 24(1515): 2 561-2 574

[6] Bezdeck J C, Ehrlich R, Full W (1984) FCM:Fuzzy c-means algorithm[J]. *Computers and Geoscience*, 23:16-20

[7] Dave R N (1996) Validating fuzzy partitions obtained through c-shells clustering[J]. *Pattern Recognition Letters*, 17(6): 613-623

[8] Vazirgiannis M, Halkidi M, Gunopulos D (2003) Uncertainty handling and quality assessment in data mining[M]. London, Hong Kong: Springer-Verlag

[9] Pal N R, Bezdek J C (1995) On cluster validity for the fuzzy c-means model[J]. *IEEE Transactions on Fuzzy Systems*, 3(3): 370-379

[10] Great Basin Center(2007) Nevada-Utah mines and prospects[OL]. http://www.unr.edu/Geothermal/GIS_download3. htm#RRvalNevada_Faults